BUNDESREPUBLIK DEUTSCHLAND

Prioritätsbescheinigung über die Einreichung einer Patentanmeldung

Aktenzeichen: 102 40 443.7

Anmeldetag: 02. September 2002

Anmelder/Inhaber: Siemens Aktiengesellschaft, München/DE

Bezeichnung: Verfahren und Anordnung sowie Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten

IPC: G 06 F 17/00

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 25. September 2003
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Brosig

A 9161
06/00
EDV-L

Beschreibung

Verfahren und Anordnung sowie Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten

Die Erfindung betrifft eine Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten, wie beispielsweise Kunden- oder Produktdaten eines Unternehmens.

Fast jeder Vorgang in einem Unternehmen, wie jeder Kontakt des Unternehmens mit einem Kunden oder jeder logistische Vorgang innerhalb eines Unternehmens, beginnend bei einer Bestellung eines Produkts bis hin zu einer Auslieferung des fertigen Produkts, wird heute elektronisch unterstützt durchgeführt bzw. kontrolliert und gesteuert.

Dabei werden systematisch Daten, beispielsweise Kundendaten oder Produktdaten, erfasst und protokolliert, die Basis für ökonomische, betriebswirtschaftliche und/oder marktstrategische Analysen sind, mit welchen die Daten in verwertbare ökonomische, betriebswirtschaftliche und/oder marktstrategische Erkenntnisse umgesetzt werden.

Wegen ihrer ökonomischen, betriebswirtschaftlichen und/oder marktstrategischen Bedeutung stellen diese Unternehmensdaten für die Unternehmen einen bedeutenden Vermögensgegenstand dar. Demzufolge unternehmen die Unternehmen große Anstrengungen bei der Erfassung und der Analyse dieser Daten.

Für die Erfassung solcher Unternehmensdaten stehen verschiedene, allgemein bekannte Systeme zur Verfügung, wie beispielsweise Customer Relationship Management Systeme (CRM) [1], Supply Chain Management Systeme (SCM) [2] oder Data Warehouses [3].

Nach der Erfassung werden die Daten meist in Datenbanken abgelegt und entsprechend strukturiert gespeichert. In der Regel werden dabei Datensätze Di=(Ai,Bi,Ci, ...) gebildet, wobei der Index i den jeweiligen Datensatz Di bezeichnet.

Jeder Datensatz Di repräsentiert ein bestimmtes Objekt aus einer Gruppe von Objekten, beispielsweise einen bestimmten Kunden aus allen erfassten Kunden eines Unternehmens oder ein bestimmtes Produkt aus einer Produktlinie eines Unternehmens.

Jeder Datensatz umfasst dabei eine vorgebbare Anzahl von Einträgen, Ai, Bi, Ci, ..., die einzelnen erfassten Daten, mit Kategorien bzw. Attributen A, B, C, ... Diese Kategorien bzw. Attribute repräsentieren Eigenschaften einer Objektgruppe, wie Alter (A), Einkommen (B), erworbenes Produkt (C), ... Die Einträge Ai, Bi, Ci, ... zu den jeweiligen Kategorien A, B, C, ... können dabei nummerischer oder semantischer Art sein.

Für die Analyse solcher Unternehmensdaten werden statistische Verfahren, sogenannte Data Mining Verfahren [4], [10], [11], [12], verwendet. Viele dieser Data Mining Verfahren bauen dabei auf einem statischen Framework auf, d.h. sie sind in einer statistischen Sprache formuliert.

Ein hinlänglich bekanntes und häufig eingesetztes Data Mining Verfahren ist ein sogenannter Entscheidungsbaum [5].

Weitere bekannte und verwendete Data Mining Verfahren sind sogenannte Clustering Verfahren [6] oder Assoziationsregeln (Association Rules) [9].

Nachteilig bei vielen der bekannten und genannten Analyseverfahren ist, dass sie bei der Analyse großer Datenmengen nur unzureichend anwendbar sind. In der Regel ist dort nämlich ein einmaliger oder mehrmaliger Zugriff auf den gesamten, zu

analysierenden Datenbestand, welcher beispielsweise in einer Datenbank gespeichert ist, notwendig.

Bei großen Datenmengen führt dies zu langen Zugriffszeiten, zu langen Rechen- und Antwortzeiten und bedingt dadurch eine schlechte Performanz. Weiter ist auch eine hohe Rechenleistung bzw. Rechenkapazität von Nöten.

Aus [7] ist eine Ermittlung eines gemeinsamen Wahrscheinlichkeitsmodells P(A, B, C, ..., X) für eine Datenstruktur (A, B, C, ...) basierend auf einer versteckten Variable X bekannt.

Aus [8] ist eine Ermittlung eines gemeinsamen Wahrscheinlichkeitsmodells P(A, B, C, ...) für eine Datenstruktur (A, B, C, ...) basierend auf ein Strukturlernen bekannt.

Der Erfindung liegt die Aufgabe zugrunde, ein Analyseverfahren zur Analyse strukturierter Nutzdaten anzugeben, welches auch bei großen Nutzdatenmengen anwendbar ist und auch dort eine hohe Performanz aufweist.

Diese Aufgabe wird durch das Verfahren und die Anordnung sowie durch das Computerprogramm mit Programmcode-Mitteln und das Computerprogramm-Produkt zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten mit den Merkmalen gemäß dem jeweiligen unabhängigen Patentanspruch gelöst.

Bei dem Verfahren zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten wird zuerst ein gemeinsames statistisches Wahrscheinlichkeitsmodell für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelt.
Anschließend werden die gemäß der Datenbankstruktur strukturierten Nutzdaten unter Verwendung eines statistischen Analyseverfahrens analysiert, wobei das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird, nicht wie üblich unmittelbar auf die Ausgangsdaten.

Die Anordnung zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten weist auf:

- eine Modellierungseinheit, mit welcher ein gemeinsames statistisches Wahrscheinlichkeitsmodell für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelbar ist, sowie
- eine Analyseeinheit, mit welcher die gemäß der Datenbankstruktur strukturierten Nutzdaten unter Verwendung eines statistischen Analyseverfahrens derart analysierbar sind, dass das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

Anschaulich gesehen basiert die Erfindung auf einer zweistufigen Vorgehensweise.

Auszugehen ist zunächst von vorgebbaren, gemäß einer Datenbankstruktur strukturierten Nutzdaten. Dabei unter einer derartigen datenbankgemäßen Strukturierung zu verstehen, dass den Nutzdaten eine übergeordnete feste Struktur zugrunde liegt, beispielsweise jeweils gleich strukturierte Datensätze (Ai, Bi, Ci, ...) mit gleichen Eintragskategorien A, B, C, ... Derartige Strukturen sind allgemein bekannt.

Aus diesen zu analysierenden, gemäß einer Datenbankstruktur strukturierten Nutzdaten wird ein gemeinsames, zweckunabhängiges Wahrscheinlichkeitsmodell, wie beispielsweise in [7], [8] beschrieben, gebildet.

Dieses stellt ein allgemeines, vollständiges und genaues Abbild einer Statistik der Datenstruktur der strukturierten Nutzdaten dar („Analytisches Datenbank-Abbild"). Ferner ist es eine hochkomprimierte Form eines Wissens über die Nutzdaten.

Das allgemeine Abbild kann dann nachfolgend als Grundlage für die Analyse durch die statistischen Verfahren verwendet werden. Diese greifen dann nicht mehr auf den gesamten Nutzdatenbestand bzw. auf die einzelnen Nutzdaten zu, sondern nutzen das erstellte statistische Abbild, d.h. das gemeinsame Wahrscheinlichkeitsmodell, für die Analyse.

Dadurch können Zugriffs-, Rechen- und Antwortzeiten bei der Analyse reduziert und damit die Performanz gesteigert werden.

Das erfindungsgemäße Computerprogramm mit Programmcode-Mitteln ist eingerichtet, um alle Schritte gemäß dem erfindungsgemäßen Analyseverfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

Das Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln ist eingerichtet, um alle Schritte gemäß dem erfindungsgemäßen Analyseverfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

Die Anordnung sowie das Computerprogramm mit Programmcode-Mitteln, eingerichtet um alle Schritte gemäß dem erfinderischen Analyseverfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird, sowie das Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln, eingerichtet um alle Schritte gemäß dem erfinderischen Analyseverfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird, sind insbesondere geeignet zur Durchführung des erfindungsgemäßen Analyseverfahrens oder einer seiner nachfolgend erläuterten Weiterbildungen.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die im weiteren beschriebenen Weiterbildungen beziehen sich sowohl auf die Verfahren als auch auf die Anordnung.

Die Erfindung und die im weiteren beschriebenen Weiterbildungen können sowohl in Software als auch in Hardware, beispielsweise unter Verwendung einer speziellen elektrischen Schaltung, realisiert werden.

Ferner ist eine Realisierung der Erfindung oder einer im weiteren beschriebenen Weiterbildung möglich durch ein computerlesbares Speichermedium, auf welchem das Computerprogramm mit Programmcode-Mitteln gespeichert ist, welches die Erfindung oder Weiterbildung ausführt.

Auch kann die Erfindung oder jede im weiteren beschriebene Weiterbildung durch ein Computerprogrammerzeugnis realisiert sein, welches ein Speichermedium aufweist, auf welchem das Computerprogramm mit Programmcode-Mitteln gespeichert ist, welches die Erfindung oder Weiterbildung ausführt.

In einer Weiterbildung werden in Nutzdatensätzen strukturierte Nutzdaten verwendet, beispielsweise Nutzdatensätze aus einer Datenbank. Dabei repräsentiert jeder Nutzdatensatz ein bestimmtes Objekt aus einer Gruppe von Objekten. Die dem jeweiligen Nutzdatensatz zugehörigen Nutzdaten beschreiben dabei Eigenschaften des jeweiligen Objekts.

Für die Ermittlung des gemeinsamen statistischen Wahrscheinlichkeitsmodell können statistische Verfahren basierend auf einer versteckten Variable [7] oder Verfahren basierend auf ein Strukturlernen [8] verwendet werden. Auch eine Kombination beider Verfahren ist möglich.

Ferner ist es zweckmäßig, dass das statistische Analyseverfahren derart auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird, dass eine gemeinsame Wahrscheinlichkeit als Eingangsgröße für das statistische Analysever-

fahren verwendet wird. Die gemeinsame Wahrscheinlichkeit ergibt sich unmittelbar aus dem gemeinsamen Wahrscheinlichkeitsmodell. Dadurch lassen sich unnötige Zwischenschritte vermeiden, die Rechenzeit kosten und Antwortzeiten verlängern.

Als statistisches Analyseverfahren kann ein Verfahren auf Basis eines Data Mining Verfahrens [4], [10], [11], [12] verwendet wird, beispielsweise ein Clustering Verfahren [5] oder ein Entscheidungsbaum [6] oder Assoziationsregeln [9].

Bei der Analyse unter Verwendung des statistischen Analyseverfahrens ist es möglich, Abhängigkeiten zwischen den Nutzdaten und/oder deren Signifikanzen basierend auf einem statistischen Test zu ermitteln. Dies kann wegen der hochkomprimierten Form der Nutzdaten, d.h. des gemeinsamen Wahrscheinlichkeitsmodells, interaktiv und sehr effizient erfolgen.

Ferner ist es sinnvoll, die Ermittlung des gemeinsamen statistischen Wahrscheinlichkeitsmodells und die Analyse des gemeinsamen statistischen Wahrscheinlichkeitsmodell durch das statistische Analyseverfahren zeit- und ortsverschieden durchzuführen.

So kann beispielsweise das Analytische Datenbank-Abbild, d.h. das gemeinsame Wahrscheinlichkeitsmodell, in vorgebbaren zeitlichen Intervallen, wie täglich oder wöchentlich, neu gebildet werden. Die Bildung kann nachts oder am Wochenende erfolgen. Das vollständige Analytische-Datenbank-Abbild steht dann bei Bedarf zur Verfügung, um Analysen erheblich zu beschleunigen.

Die Nutzdaten können aus verschiedenen Datenquellen bezogen werden. Am einfachsten ist der Bezug der Nutzdaten aus einer Datenbank, in welcher die Nutzdaten gespeichert sind und von welcher sie ausgelesen werden.

Die Erfindung ist wegen der durch sie erreichbaren Performanz bei der Analyse von Daten insbesondere dort geeignet, wo große Datenmengen zu verarbeiten bzw. zu analysieren sind, wie im Bereich eines Customer Relationship Management (CRM) [1] oder eines Supply Chain Management [2] oder eines Data Warehouse (DW) [3].

Im Bereich CMR kann eine Weiterbildung beispielsweise dazu eingesetzt werden, um Kundendaten zu analysieren. In diesem Fall ist das Objekt ein Kunde, welcher durch mindestens zwei der folgenden Eigenschaften, Alter, Einkommen, erworbenes Produkt, Datum des Erwerbs, Häufigkeit von Käufen, beschrieben wird. Dadurch lassen sich für Marketingabteilungen eminent wichtige Fragestellungen lösen, wie ein Kundenverhalten bestimmter Kundengruppen. Basierend darauf lassen sich gezielter Zielgruppen bei einer Akquisition von Kunden bestimmen, für bestimmte Produkte und Marketingkampagnen sinnvoller Kundengruppen auswählen und Kunden allgemein vorausschauender bedienen.

Ein Ausführungsbeispiel der Erfindung ist in Figuren dargestellt und wird im weiteren erläutert.

Es zeigen

Figur 1 Skizze, die schematisch eine Funktionsweise eines Analysesystems zur Analyse von Kundendaten gemäß einem Ausführungsbeispiel zeigt;

Figuren 2a bis g Skizzen, die Analyseergebnisse eines Analysesystems zur Analyse von Kundendaten gemäß einem Ausführungsbeispiel zeigen.

Ausführungsbeispiel:

Analysesystem zur Analyse eines Kundenverhaltens bei einer Bank basierend auf einem Customer Relationship Management System

Gegenstand des Ausführungsbeispiels ist ein Analysesystem zur Analyse von Kundendaten einer Bank.

Vorwegzuschicken ist, dass das im Folgende beschriebene Analysesystem nicht nur bei Banken, sondern auch bei beliebigen Unternehmen zur Analyse von entsprechenden Unternehmensdaten einsetzbar ist, wie beispielsweise bei Warenhäuser oder produzierenden Unternehmen.

Funktionsweise des Analysesystems (Fig.1)

Fig.1 zeigt schematisch die Funktionsweise 100 des Analysesystems zur Analyse der Bankkundendaten 110.

Die Funktionsweise 100 teilt sich auf in eine Wissensgewinnung 101 und eine Umsetzung des Wissens in eine intelligente Bedienung der Bankkunden 102.

Große und damit schwer handhabbare Mengen von Kundendaten 110 werden zunächst zu einem statistischen Modell 112, einem gemeinsamen Wahrscheinlichkeitsmodell, des Kundenverhaltens kondensiert 111.

Das verwendete gemeinsame Wahrscheinlichkeitsmodell 112 ist eines auf der Basis einer versteckten Variablen. Grundlagen dazu sind in [7] beschrieben.

Anzumerken ist, dass auch andere Arten von gemeinsamen Wahrscheinlichkeitsmodellen verwendet werden können, wie beispielsweise solche auf der Basis von Strukturlernen [8].

An Hand des gemeinsamen Wahrscheinlichkeitsmodells 112 lassen sich Eigenschaften der Kunden und insbesondere deren Verhalten über die Zeit sehr viel effizienter und flexibler explorieren als an Hand der Ausgangsdaten.

Dazu werden statistische Verfahren 120, im allgemeinen Data Mining Verfahren und hier in diesem Fall ein Entscheidungsbaum, verwendet, welche bzw. welcher auf das statistische Modell aufsetzen bzw. aufsetzt.

Anzumerken ist, dass auch andere Data Mining Verfahren verwendet werden können, wie beispielsweise Clustering Verfahren oder Assoziations-Regeln.

Grundlagen von Data Mining Verfahren sind in [4], [10], [11], [12], eines Entscheidungsbaums in [6] und von Clustering Verfahren in [5] beschrieben.

Ermöglicht wird die Kopplung dadurch, dass die Data Mining Verfahren bzw. der Entscheidungsbaum 120 auf einem statistischen Framework aufbauen bzw. aufbaut und damit die gleiche statistischen Begriffe bzw. die gleiche statistische Sprache wie das gemeinsame Wahrscheinlichkeitsmodell 112 benutzt.

Wichtige Fragestellungen (vgl. Figuren 2) können anhand des Entscheidungsbaums 120 im Rückgriff auf das gemeinsame Wahrscheinlichkeitsmodell 112 interaktiv beantwortet werden 140.

Damit ist nicht nur eine quantitative (wie viel Kunden?) sondern auch eine qualitative Sicht auf die Kunden (welche Sorte von Kunden) möglich, z.B.:

- Wie viele und welche Qualität von Kunden kommen über welche Partnerschaften oder Kampagnen? Wie effizient sind meine Werbemaßnahmen?
- Welche Kundenklassen mit welchen Präferenzen und Bedürfnissen gibt es? Wie und wann lassen sich diese Bedürfnisse am besten befriedigen?

Ergebnisse der Fragestellungen lassen sich weiterführend umsetzen 121 in eine intelligente Bedienung der Kunden 130.

Kundendaten ((Fig.1, 110)

Die Kundendaten 110 bei dem Analysesystem werden im Rahmen eines Customer Relationship Management (CRM) 150 erhoben.

Grundlagen eines CRM sind in [1] beschrieben.

Bei dem CRM 150 werden große Mengen an Daten 110 über die Bankkunden aus allen Vertriebskanälen der Bank, wie direkte Kontakte, Web, Call Center, erfasst und gespeichert.

Erfasst und gespeichert werden für die Kunden jeweils (sogenannte Attribute A, B, C, ...) :

- die erworbenen Bankprodukte A in der jeweiligen zeitlichen Reihenfolge (A1, A2, A3, ...),
- ein zeitlicher Kaufabstand B zwischen den Erwerbszeitpunkten der erworbenen Bankprodukten (B1-2, B2-3, B3-4, ...),
- ein Geburtsdatum (C),
- ein Einkommen (D),
- eine Adresse (E),
- der letzter Bankbesuch (F),
- die letze Kontobewegung (G).

Die Speicherung erfolgt in einer Datenbank in Form von kundenspezifischen Datensätzen Di(A1, A2, ..., B1-2, B2-3, ..., C, D, ...), wobei der Index i den jeweiligen Bankkunden i kennzeichnet.

Gemeinsames Wahrscheinlichkeitsmodell (Fig.1, 112)

Das Wissen über die Bankkunden, das in diesen Daten 110 verborgen liegt, wird dann zu einem Modell, dem gemeinsamen Wahrscheinlichkeitsmodell 112, kondensiert.

Das verwendete gemeinsame Wahrscheinlichkeitsmodell 112 ist eines auf der Basis einer versteckten Variablen X. Grundlagen dazu sind in [7] beschrieben.

Geschrieben wird das gemeinsame Wahrscheinlichkeitsmodell 112 basierend auf der versteckten Variablen X als P(A,B, C, ..., X) für alle Attribute (A, B, C, ...).

Ein solches statistischen Abbild von Daten stellt eine hochkomprimierte Form eines Wissens über Kunden dar und kann genutzt werden, um effizient und interaktiv Abhängigkeiten zu explorieren 120, 140.

An Hand des hier erstellten gemeinsamen Wahrscheinlichkeitsmodells 112 läßt sich nun das Wissen über die Kunden schnell über effizient abgreifen, insbesondere lassen sich Verhaltensweisen der Kunden einfach und flexibel studieren, lassen sich typische Verhaltensmuster und Entwicklungszyklen von Kunden effizient und intuitiv analysieren, lassen sich typische Kundensegmente und deren Präferenzen sicher und eindeutig bestimmen und erkennen 120, 140.

Ferner liefert das gemeinsame Wahrscheinlichkeitsmodell 112 über die beschriebene Analysefunktion hinaus schnell abrufbare Prognosen über weiter zu erwartendes Verhalten und aktuelle Bedürfnisse eines Kunden. Die Prognosen können weiter dazu genutzt werden, Kunden vorausschauend und gezielt zu bedienen und proaktive, persönliche Angebote zu unterbreiten 130.

Aufsatz eines Entscheidungsbaums auf das gemeinsame Wahrscheinlichkeitsmodell (Fig.1, 120)

In weiterer Verwendung des gemeinsamen Wahrscheinlichkeitsmodells 112 wird der Entscheidungsbaum [6] auf das statistische Modell 112, das gemeinsame Wahrscheinlichkeitsmodell 112, aufgesetzt 120.

Damit lassen sich beliebige Randverteilungen, wie die für einen ersten Split des Entscheidungsbaums, nämlich P(A,X), P(B,X), P(C,X), ..., und auch für alle weiteren Splits des Entscheidungsbaums ermitteln.

Weiter lassen sich auch alle Grundwahrscheinlichkeitsverteilungen bzw. Grundwahrscheinlichkeiten P(A), P(B), ... und beliebige bedingte Wahrscheinlichkeiten bzw. Wahrscheinlichkeitsverteilungen P(B|A), P(C|A), P(C|B), ... ermitteln.

Aus der gemeinsamen Verteilung P(A,B, C, ..., X) basierend auf der versteckten (oder latenten) Variable X geht zunächst die gemeinsame Verteilung P(A,B, C, ...) über alle Attribute der Kunden durch Summation über die versteckte Variable X hervor.

Strukturlernen liefert hier unmittelbar eine gemeinsame Verteilung P(A,B, C, ...).

Aus der gemeinsamen Verteilung lassen sich dann beliebige ein-dimensionale Randverteilungen (Marginale) P(A), P(B), ..., niedrig-dimensionalere Verteilungen P(A,B), P(B,C), ... und beliebige bedingte Wahrscheinlichkeiten (ein- oder mehrdimensionale) P(B|A), P(C|A), P(A,C|B), ... ableiten.

Dies erfolgt im Rahmen eines Inferenzprozesses, wie in [13] beschrieben.

Dabei wird nach [13] die Struktur der Modelle, beispielsweise welche mit einer vorgegebenen versteckten Variable oder welche, die durch Strukturlernen erzeugt wurden, oder eine Kombination der Vorgenannten, genutzt, um notwendigen Summen über die gemeinsame Verteilung effizient zu berechnen.

Entscheidungsbäume werden zumeist nach einem bekannten CHAID oder einem bekannten CART Verfahren aufgebaut.

Im Allgemeinen benötigt man zum Aufbau eines Entscheidungsbaums mit einer Zielvariablen (oder abhängigen Variablen) A für den sogenannten ersten Split zunächst alle paarweisen Verteilungen P(A,B), P(B,C), P(A,D), ...

Eine Selektion einer Variablen aus der Menge der Variablen B, C, D, ..., für den ersten Split erfolgt dann bei fast allen bekannten Verfahren basierend auf einem statistischen Kriterium (einem statistischen Test und Signifikanzkriterien) basierend auf den paarweisen Verteilungen P(A,B), P(B,C), P(A,D), ... und einer bekannten Anzahl an Daten.

Wurde beispielsweise für den ersten Split die Variable D mit den beiden Werten d1 und d2 gewählt., so benötigt man für den zweiten Split bedingte, paarweise Verteilungen der Form P(A,B|d1), P(A,B|d2), P(A,C|d1), P(A,C|d2), ...

Die notwendigen Wahrscheinlichkeiten oder Verteilungen für den Aufbau des Entscheidungsbaums (bzw. als Grundlagen für die notwendigen statistischen Tests) können (wie üblich) aus den Daten oder auch aus einem möglichst genauen, im Obigen beschrieben Wahrscheinlichkeitsmodell (Inferenzprozess) ermittelt werden.

Interaktive Analysen (Fig.1, 140, Fig.2a bis 2g)

Fig. 2a bis 2g zeigen exemplarisch einige der möglichen interaktiven Analysen 140, welche mit dem Entscheidungsbaum 120 im Rückgriff auf das gemeinsame Wahrscheinlichkeitsmodell 112 durchgeführt werden können.

Fig.2a zeigt Wahrscheinlichkeitsverteilungen P(A1), P(A2) P(A3) P(A4), P(A5), P(B1-2), P(B2-3), P(B3-4) und P(C) und P(D). Besondere gekennzeichnet ist P(A1= „Giro/Gehalts-Konto)=56,125%.

Fig.2b zeigt nun bedingte Wahrscheinlichkeitsverteilungen unter der Bedingung A1= „Giro/Gehalts-Konto", nämlich P(A2|A1= „Giro/Gehalts-Konto"), P(A3|A1= „Giro/Gehalts-Konto"), P(A4|A1= „Giro/Gehalts-Konto"), P(A5|A1= „Giro/Gehalts-Konto"), P(B1-2|A1= „Giro/Gehalts-Konto"), P(B2-3|A1= „Gi-

ro/Gehalts-Konto"), P(B3-4|A1= „Giro/Gehalts-Konto") und P(C|A1= „Giro/Gehalts-Konto") und P(D|A1= „Giro/Gehalts-Konto"). Besonders gekennzeichnet sind P(A2= „Versicherungspodukt|A1= „Giro/Gehalts-Konto)=29% und P(A2= „Sparen/Geldanlage|A1= „Giro/Gehalts-Konto)=50%.

Fig.2c zeigt nun bedingte Wahrscheinlichkeitsverteilungen unter den Bedingungen A1= „Giro/Gehalts-Konto" und A2= „Versicherungsprodukt", nämlich P(A3|A1= „Giro/Gehalts-Konto", A2= „Versicherungsprodukt"), P(A4|A1= „Giro/Gehalts-Konto", A2= „Versicherungsprodukt"), P(A5|A1= „Giro/Gehalts-Konto", A2= „Versicherungsprodukt"), ... Besonders gekennzeichnet ist hier P(B1-2= „Kaufabstand zwischen erstem und zweitem Produkt größer 3 Jahre|A1= „Giro/Gehalts-Konto",A2= „Versicherungsprodukt)=85%.

Fig.2d zeigt weitere bedingte Wahrscheinlichkeitsverteilungen unter den Bedingungen A1= „Giro/Gehalts-Konto" und A2= „Sparen/Geldanlage", nämlich P(A3|A1= „Giro/Gehalts-Konto", A2= „Sparen/Geldanlage"), P(A4|A1= „Giro/Gehalts-Konto", A2= „Sparen/Geldanlage"), P(A5|A1= „Giro/Gehalts-Konto", A2= „Sparen/Geldanlage"), ... Besonders gekennzeichnet sind hier die Wahrscheinlichkeitsverteilungen P(B1-2|A1= „Giro/Gehalts-Konto",A2= „Sparen/Geldanlage").

Fig.2e zeigt die Wahrscheinlichkeitsverteilungen P(A1), P(A2) P(A3) P(A4), P(A5), P(B1-2), P(B2-3), P(B3-4) und P(C) und P(D). Besondere gekennzeichnet ist P(A1= „Giro/Gehalts-Konto)=56,125%. Desweiteren zeigt Fig.2e die Wahrscheinlichkeitsverteilung der versteckten Variable X, bezeichnet hier als Segmente, nämlich P(Segmente). Besonders gekennzeichnet st P(Segmente=4)= 34%, was zeigt, dass 34% aller erfassten Bankkunden in das Segment 4 fallen.

Figuren 2f und 2g zeigen wiederum die bedingte Wahrscheinlichkeitsverteilungen, einmal unter der Bedingung Segmente=4

(Fig.2f) und das andere Mal unter der Bedingung C= Geburtsdatum zwischen 980 und 1990 (Fig.2g).

Im Rahmen dieses Dokuments sind folgende Veröffentlichungen zitiert:

[1] Customer Relationship Management System, erhältlich am 31.08.2002 unter: http://www.crm-expo.com/.

[2] Supply Chain Management System, erhältlich am 31.06.2002 unter: http://www.sap-ag.de/germany/solutions/scm/.

[3] Data Warehouse, erhältlich am 31.08.2002 unter: http://www.data-warehouse-systeme.de/.

[4] Heckermann, D., „Bayesian Networks for Data Mining", Data Mining and Knowledge Discovery, Seiten 79 bis 119, 1997.

[5] Kass, G., „An exploratory technique for investigating large quantities of categorical data", Applied Statistics, 29:2, Seiten 119 bis 117, 1980.

[6] Bezdek, J.C., Pal, S.K., „Fuzzy Models for Pattern Recognition", IEEE Press, 1992.

[7] Everitt, B. S., „An Introduction to Latent Variable Models", London, Chapman and Hall, 1984.

[8] Reimar Hofmann, „Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen", Dissertation an der Technischen Universität München, Verlag: dissertation.de, ISBN:3-89825-131-4.

[9] Ashoka Savasere, Edward Omiecinski, Shamkant B. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", The VLDB Journal, Seiten 432 bis 444", 1995.

[10] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining", American Association for Artificial Intelligence, CA, 1996.

[11] Ian H. Witten, Eibe Frank, Morgan Kaufmann, Data Mining, 2000.

[12] T. Hastie, R. Tibshirani, J. H. Friedman, „The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer Series in Statistics.

[13] Jensen, V. J., "An Introduction to Bayesian Networks", UCL Press, London, 1996.

Patentansprüche

1. Verfahren zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten,
- bei dem ein gemeinsames statistisches Wahrscheinlichkeitsmodell für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelt wird,
- bei dem die gemäß der Datenbankstruktur strukturierten Nutzdaten unter Verwendung eines statistischen Analyseverfahrens analysiert werden, wobei das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

2. Verfahren nach Anspruch 1,
bei dem die gemäß der Datenbankstruktur strukturierten Nutzdaten in Nutzdatensätze strukturiert sind, welche Nutzdatensätze jeweils ein Objekt repräsentieren, wobei die Nutzdaten eines Nutzdatensatzes Eigenschaften des jeweiligen Objekts beschreiben.

3. Verfahren nach Anspruch 1 oder 2,
bei dem das gemeinsame statistische Wahrscheinlichkeitsmodell basierend auf einer versteckten Variable ermittelt wird.

4. Verfahren nach einem der Ansprüche 1 bis 3,
bei dem das gemeinsame statistische Wahrscheinlichkeitsmodell basierend auf ein Strukturlernen ermittelt wird.

5. Verfahren nach einem der vorangehenden Ansprüche,
bei dem das statistische Analyseverfahren derart auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird, dass eine gemeinsame Wahrscheinlichkeit des gemeinsamen Wahrscheinlichkeitsmodells als Eingangsgröße für das statistische Analyseverfahren verwendet wird.

6. Verfahren nach einem der vorangehenden Ansprüche,

bei dem als statistisches Analyseverfahren ein Verfahren auf Basis eines Data Mining Verfahrens verwendet wird.

7. Verfahren nach Anspruch 6,
bei dem als statistisches Analyseverfahren ein Clustering Verfahren verwendet wird.

8. Verfahren nach Anspruch 6
Bei dem als statistisches Analyseverfahren ein Verfahren bekannt unter dem Namen „Assoziationsregeln" verwendet wird.

9. Verfahren nach Anspruch 6,
bei dem als statistisches Analyseverfahren ein Entscheidungsbaum verwendet wird.

10. Verfahren nach einem der vorangehenden Ansprüche,
bei dem bei der Analyse unter Verwendung des statistischen Analyseverfahrens Abhängigkeiten zwischen den Nutzdaten ermittelt werden und/oder deren Signifikanzen basierend auf einem statistischen Test ermittelt werden.

11. Verfahren nach einem der vorangehenden Ansprüche,
bei dem die Ermittlung des gemeinsamen statistischen Wahrscheinlichkeitsmodells und die Analyse des gemeinsamen statistischen Wahrscheinlichkeitsmodell durch das statistische Analyseverfahren zeit- und ortsverschieden durchgeführt werden.

12. Verfahren nach einem der vorangehenden Ansprüche,
bei dem die Nutzdaten in einer Datenbank gespeichert werden.

13. Verfahren nach einem der Ansprüche 2 bis 12,
bei dem das Objekt ein Kunde ist, welcher durch mindestens zwei der folgenden Eigenschaften, Alter, Einkommen, erworbenes Produkt, Datum des Erwerbs, Häufigkeit von Käufen, beschrieben wird.

14. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt bei dem Data Warehouse, wobei die Nutzdaten das Data Warehouse beschreiben.

15. Verfahren nach einem der Ansprüche 1 bis 13, eingesetzt bei einem Customer Relationship Management oder einem Supply Chain Management, wobei die Nutzdaten Kundendaten oder Produktdaten sind.

16. Anordnung zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten,

- mit einer Modellierungseinheit, mit welcher ein gemeinsames statistisches Wahrscheinlichkeitsmodell für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelbar ist,
- mit einer Analyseeinheit, mit welcher die gemäß der Datenbankstruktur strukturierten Nutzdaten unter Verwendung eines statistischen Analyseverfahrens derart analysierbar sind, dass das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

17. Computerprogramm-Erzeugnis, das ein computerlesbares Speichermedium umfasst, auf dem ein Programm gespeichert ist, das es einem Computer ermöglicht, nachdem es in einen Speicher des Computers geladen worden ist, folgende Schritte durchzuführen zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten,

- ein gemeinsames statistisches Wahrscheinlichkeitsmodell wird für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelt,
- die gemäß der Datenbankstruktur strukturierten Nutzdaten werden unter Verwendung eines statistischen Analyseverfahrens analysiert, wobei das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

18. Computerlesbares Speichermedium, auf dem ein Programm gespeichert ist, das es einem Computer ermöglicht, nachdem es in einen Speicher des Computers geladen worden ist, folgende Schritte durchzuführen zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten,

- ein gemeinsames statistisches Wahrscheinlichkeitsmodell wird für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelt,
- die gemäß der Datenbankstruktur strukturierten Nutzdaten werden unter Verwendung eines statistischen Analyseverfahrens analysiert, wobei das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

19. Computerprogramm mit Programmcode-Mitteln, um alle Schritte gemäß Anspruch 1 durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

20. Computerprogramm mit Programmcode-Mitteln gemäß Anspruch 18, die auf einem computerlesbaren Datenträger gespeichert sind.

21. Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln, um alle Schritte gemäß Anspruch 1 durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

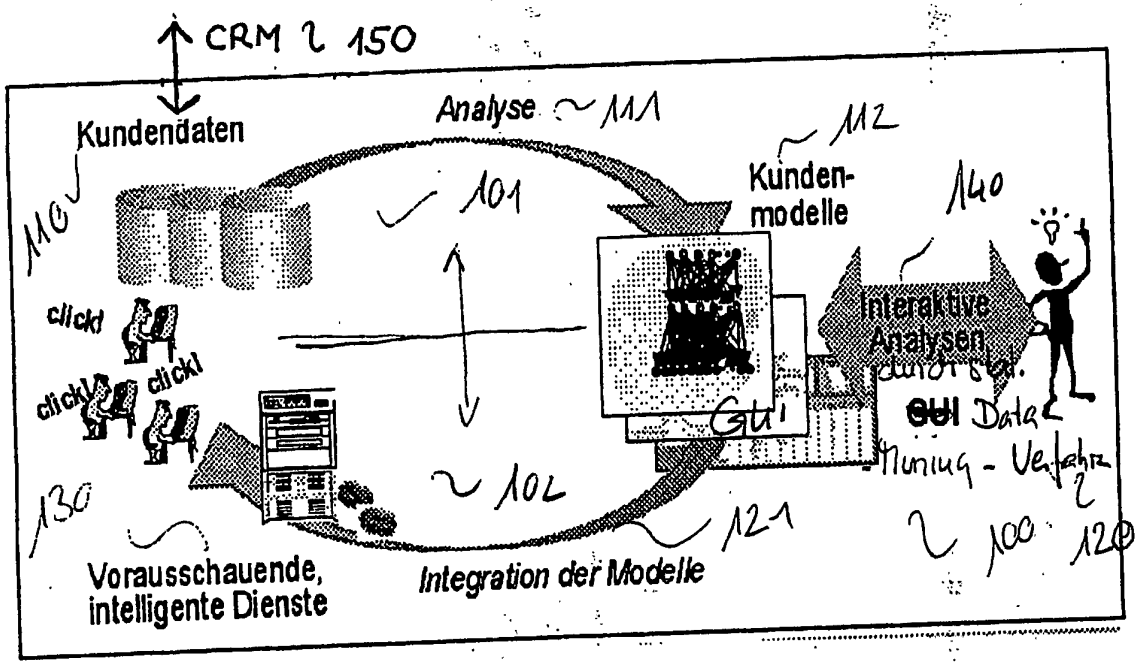
Zusammenfassung

Verfahren und Anordnung sowie Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Analyse von gemäß einer Datenbankstruktur strukturierten Nutzdaten

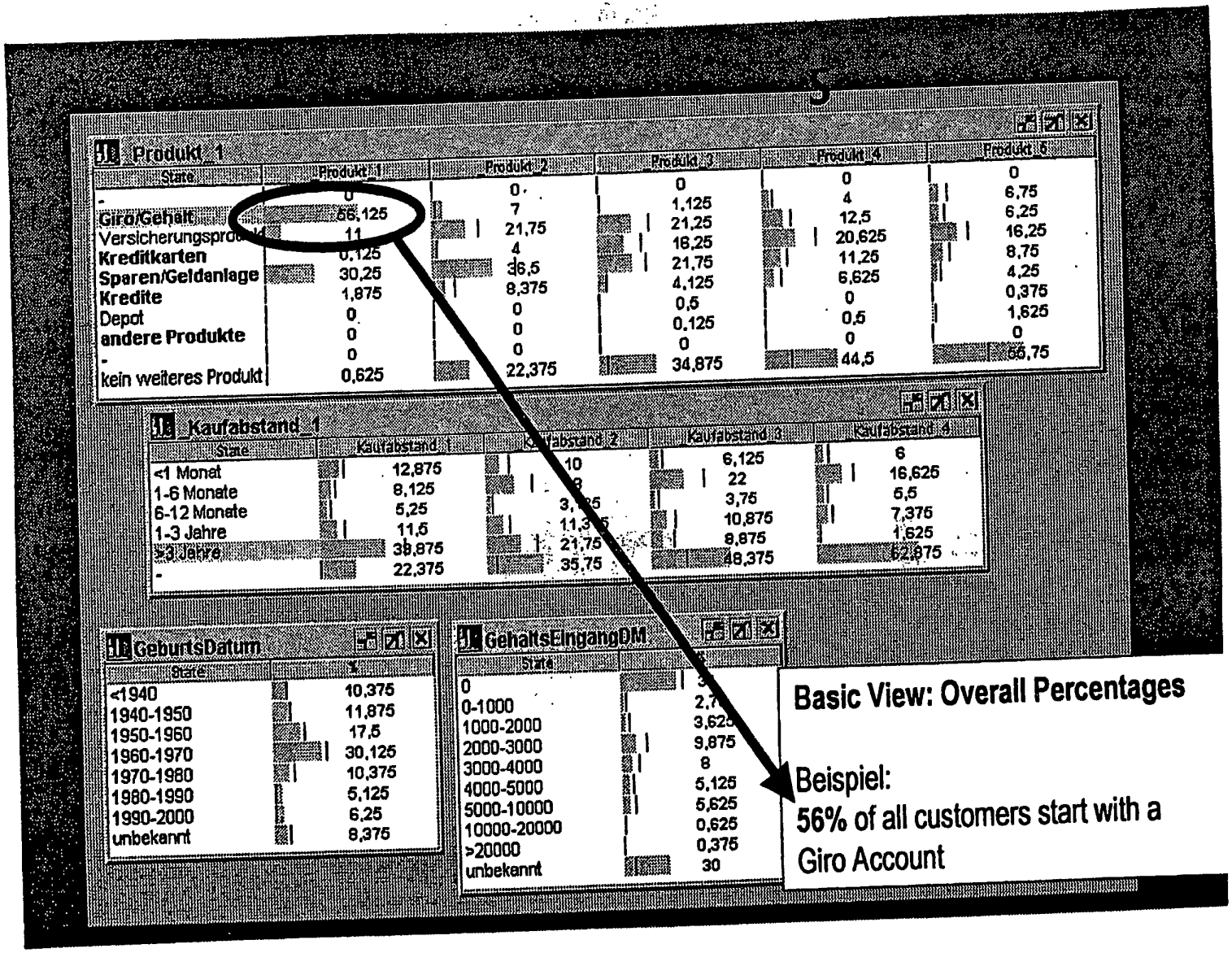
Bei der Analyse wird zuerst ein gemeinsames statistisches Wahrscheinlichkeitsmodell für die gemäß der Datenbankstruktur strukturierten Nutzdaten ermittelt. Anschließend werden die gemäß der Datenbankstruktur strukturierten Nutzdaten unter Verwendung eines statistischen Analyseverfahrens analysiert, wobei das bei der Analyse verwendete statistische Analyseverfahren auf das gemeinsame statistische Wahrscheinlichkeitsmodell angewendet wird.

Sign. Fig. 1

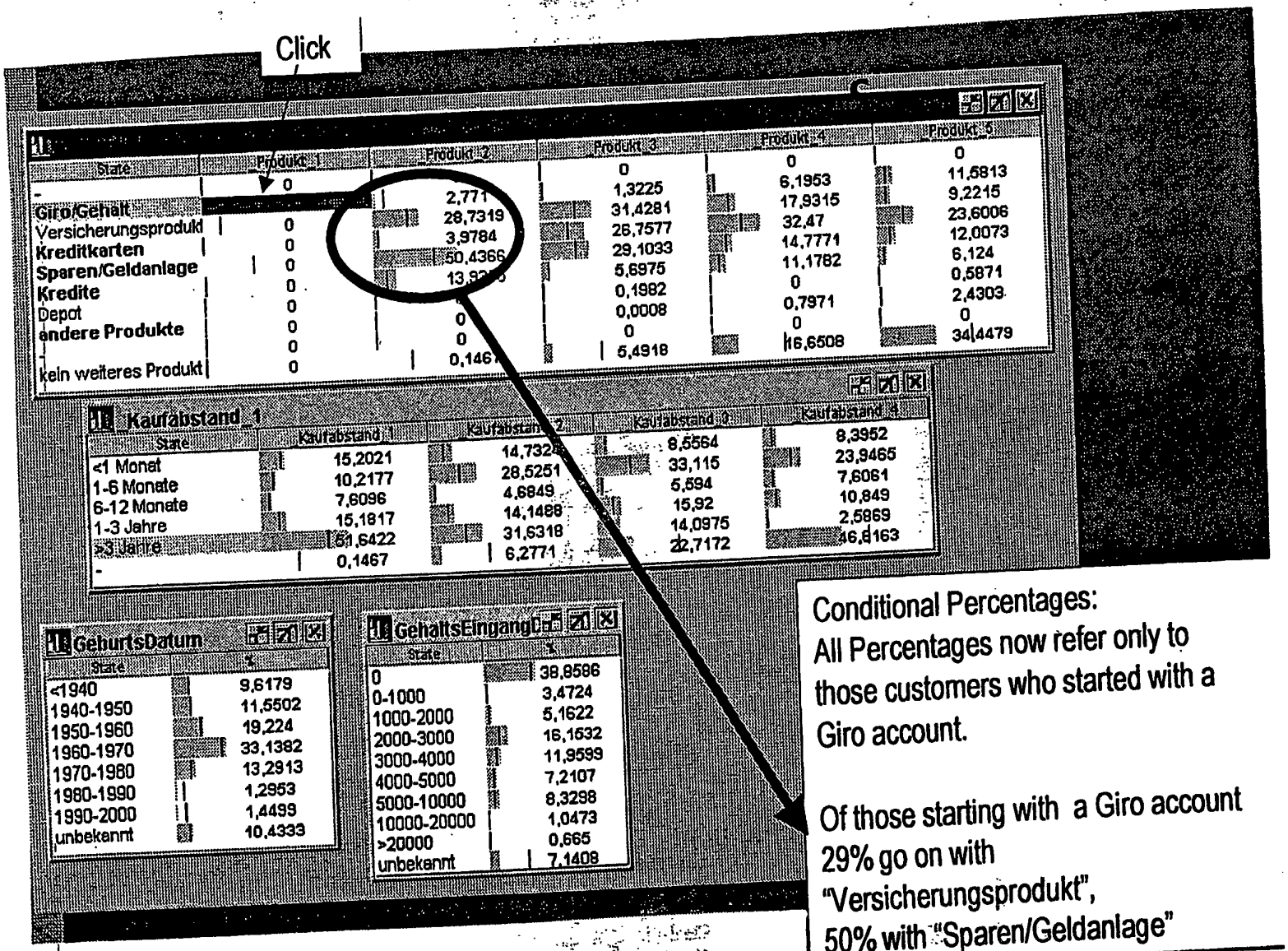
Figur 1



Figur 2a


Produkt_1
State | Produkt_1 | Produkt_2 | Produkt_3 | Produkt_4 | Produkt_5
- | 0 | 0 | 0 | 0 | 0
Giro/Gehalt | 56,125 | 7 | 1,125 | 4 | 6,75
Versicherungsprodukt | 11 | 21,75 | 21,25 | 12,5 | 6,25
Kreditkarten | 0,125 | 4 | 18,25 | 20,625 | 16,25
Sparen/Geldanlage | 30,25 | 36,5 | 21,75 | 11,25 | 8,75
Kredite | 1,875 | 8,375 | 4,125 | 6,625 | 4,25
Depot | 0 | 0 | 0,5 | 0 | 0,375
andere Produkte | 0 | 0 | 0,125 | 0,5 | 1,625
- | 0 | 0 | 0 | 0 | 0
kein weiteres Produkt | 0,625 | 22,375 | 34,875 | 44,5 | 55,75
Kaufabstand_1
State | Kaufabstand_1 | Kaufabstand_2 | Kaufabstand_3 | Kaufabstand_4
<1 Monat | 12,875 | 10 | 6,125 | 6
1-6 Monate | 8,125 | 8 | 22 | 16,625
6-12 Monate | 5,25 | 3,75 | 3,75 | 5,5
1-3 Jahre | 11,5 | 11,25 | 10,875 | 7,375
>3 Jahre | 38,875 | 21,75 | 8,875 | 1,625
- | 22,375 | 35,75 | 48,375 | 62,875
GeburtsDatum
State | %
<1940 | 10,375
1940-1950 | 11,875
1950-1960 | 17,5
1960-1970 | 30,125
1970-1980 | 10,375
1980-1990 | 5,125
1990-2000 | 6,25
unbekannt | 8,375
GehaltsEingangDM
State
0
0-1000
1000-2000
2000-3000
3000-4000
4000-5000
5000-10000
10000-20000
>20000
unbekannt
9,875
8
5,125
5,625
0,625
0,375
30
Basic View: Overall Percentages
Beispiel:
56% of all customers start with a Giro Account

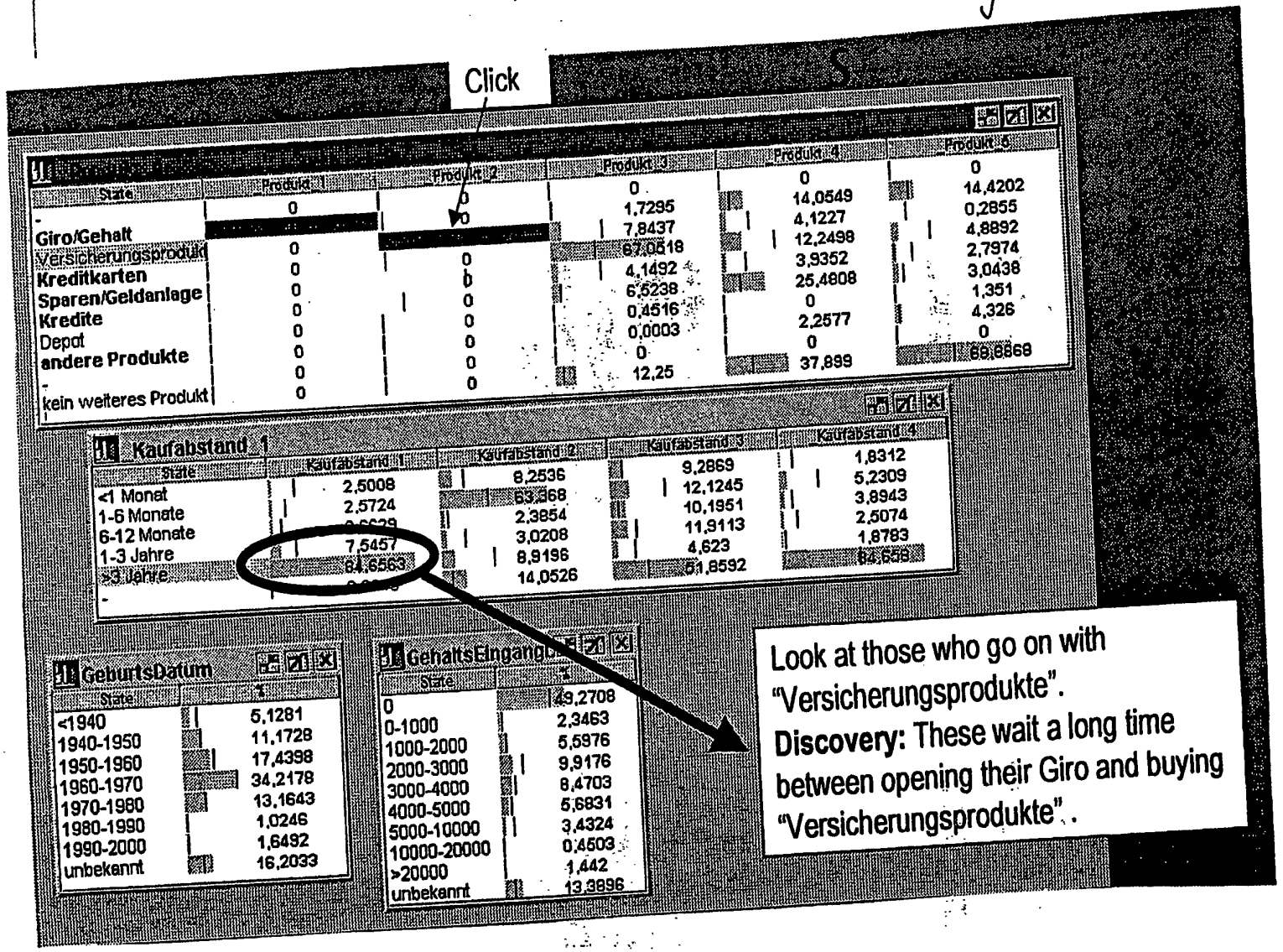
Figur 26



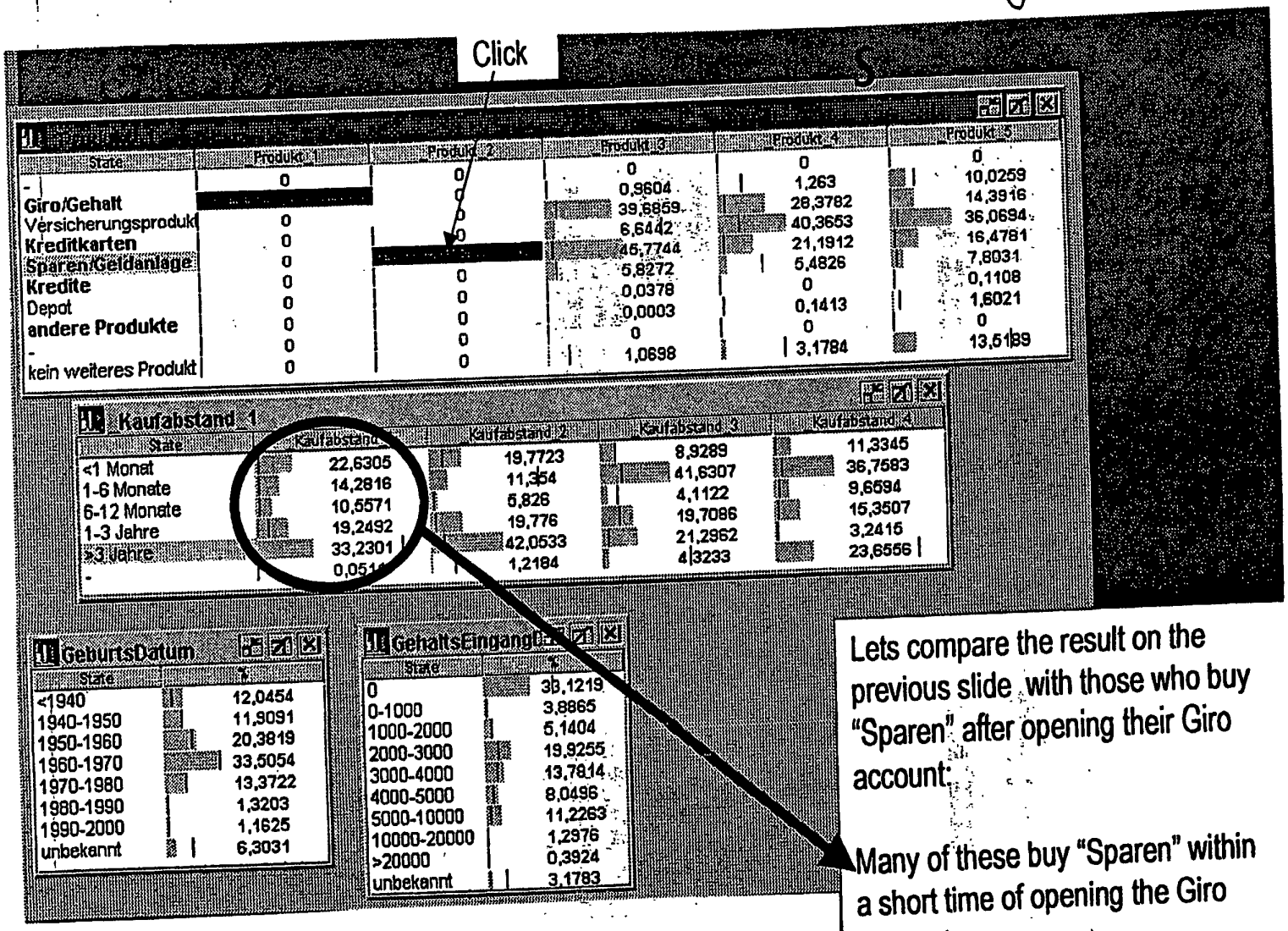
Conditional Percentages:
All Percentages now refer only to those customers who started with a Giro account.

Of those starting with a Giro account
29% go on with
"Versicherungsprodukt",
50% with "Sparen/Geldanlage"

Figur 2c


Click
State
Produkt 1
Produkt 2
Produkt 3
Produkt 4
Produkt 5
Giro/Gehalt
Versicherungsprodukt
Kreditkarten
Sparen/Geldanlage
Kredite
Depot
andere Produkte
kein weiteres Produkt
Kaufabstand 1
Kaufabstand 2
Kaufabstand 3
Kaufabstand 4
<1 Monat
1-6 Monate
6-12 Monate
1-3 Jahre
>3 Jahre
GeburtsDatum
GehaltsEingang
Look at those who go on with "Versicherungsprodukte".
Discovery: These wait a long time between opening their Giro and buying "Versicherungsprodukte".

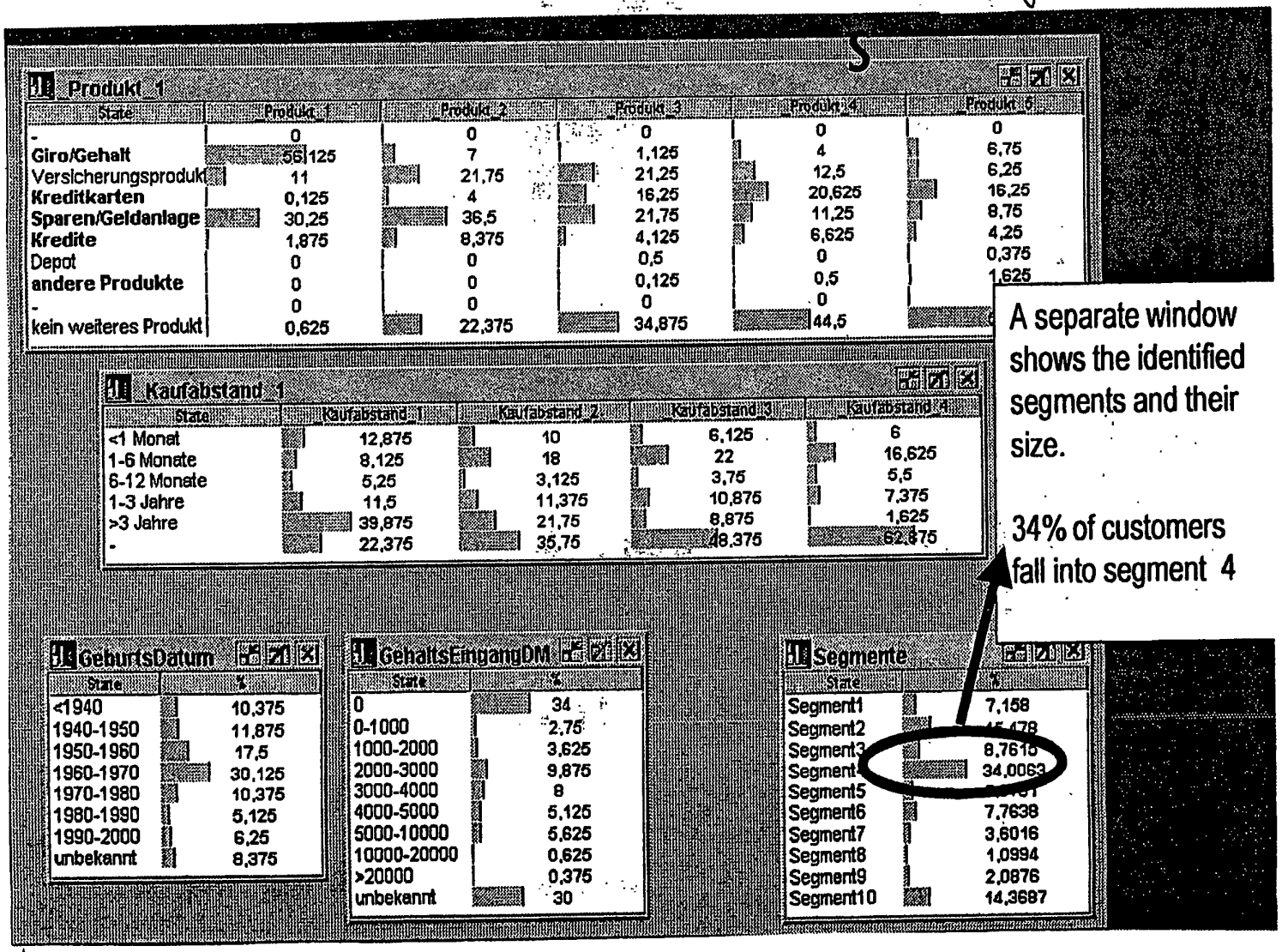
Figur 2d



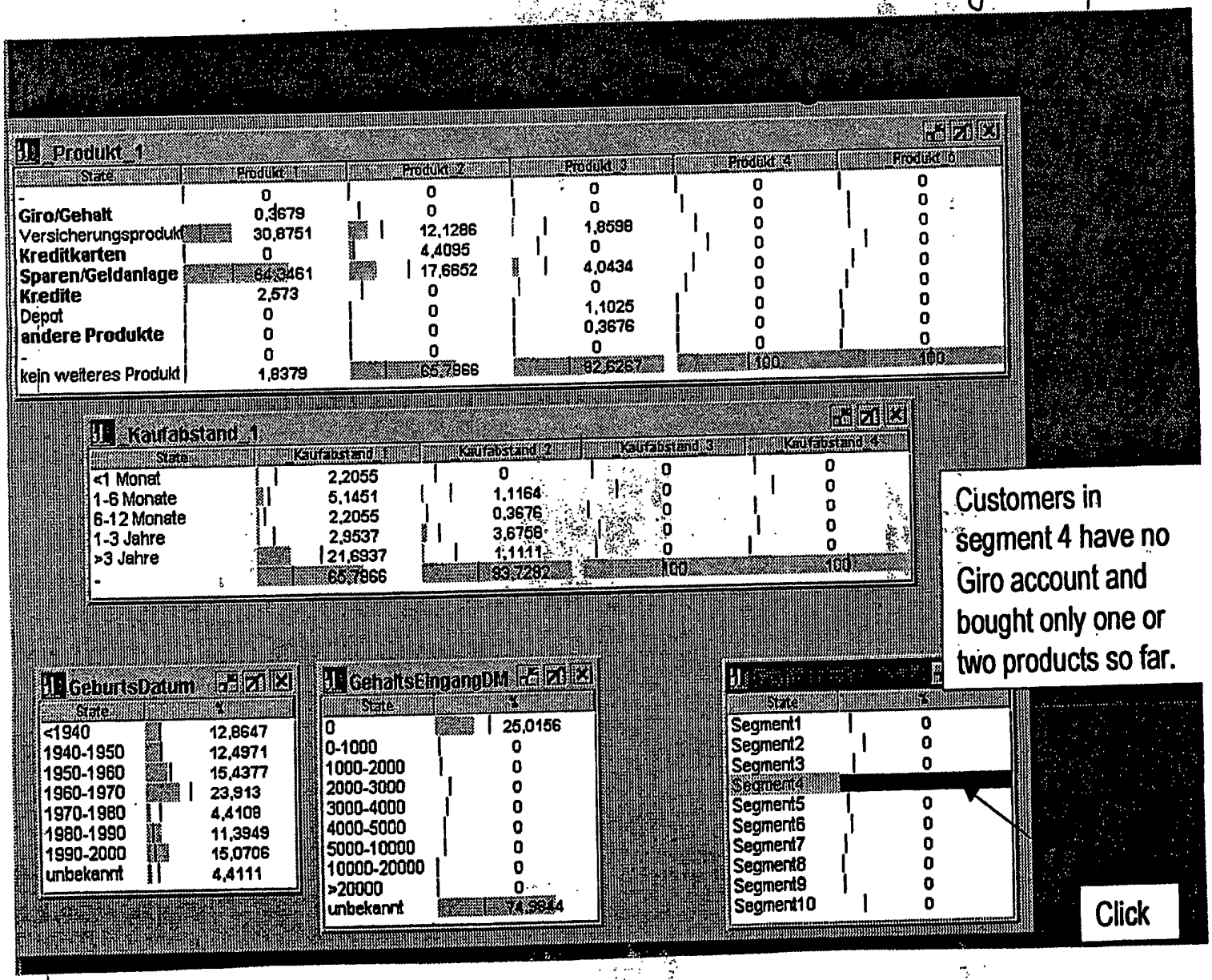
Lets compare the result on the previous slide with those who buy "Sparen" after opening their Giro account:

Many of these buy "Sparen" within a short time of opening the Giro account.

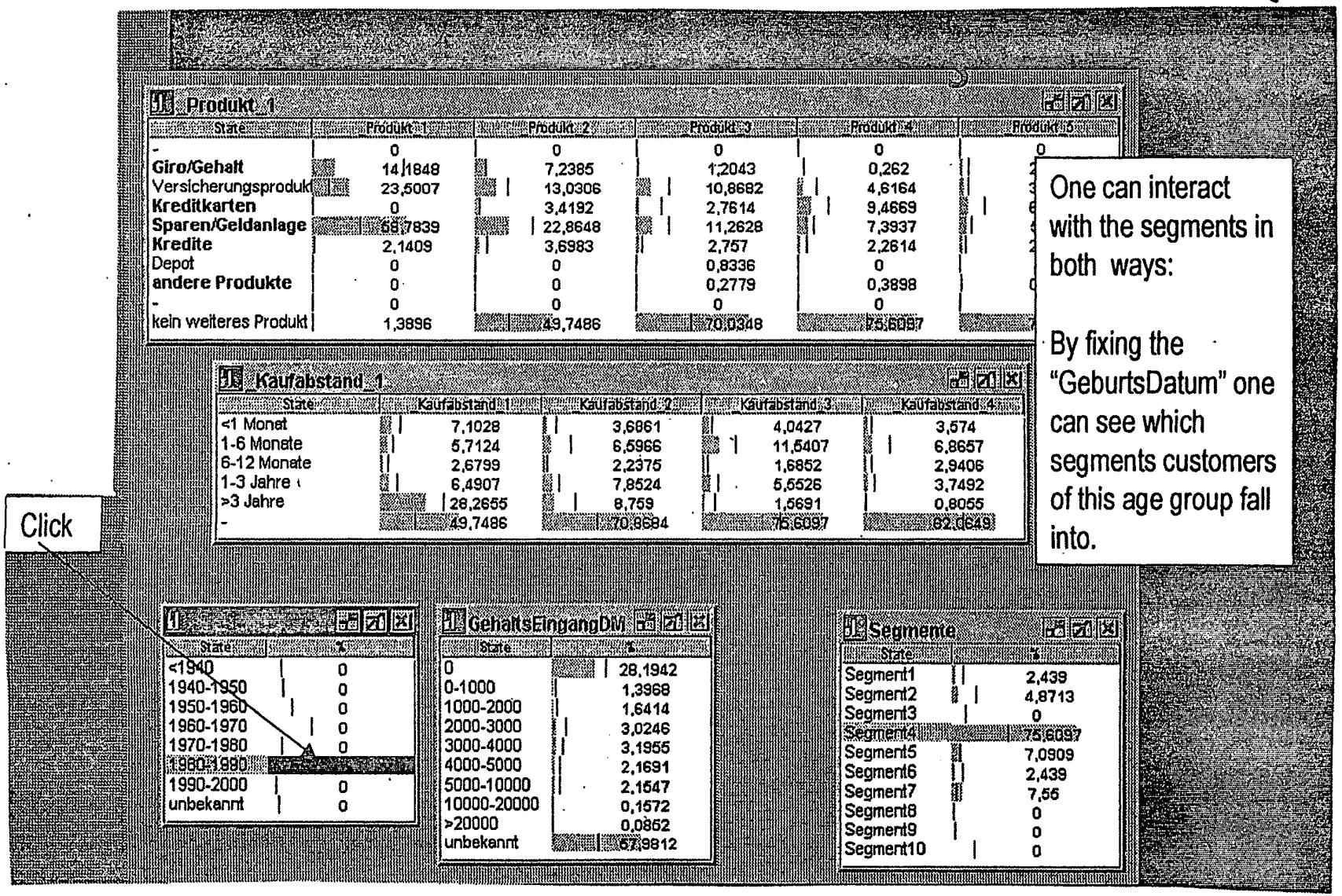
Figur 2e


Produkt_1
State | Produkt_1 | Produkt_2 | Produkt_3 | Produkt_4 | Produkt_5
- | 0 | 0 | 0 | 0 | 0
Giro/Gehalt | 56,125 | 7 | 1,125 | 4 | 6,75
Versicherungsprodukt | 11 | 21,75 | 21,25 | 12,5 | 6,25
Kreditkarten | 0,125 | 4 | 16,25 | 20,625 | 16,25
Sparen/Geldanlage | 30,25 | 36,5 | 21,75 | 11,25 | 8,75
Kredite | 1,875 | 8,375 | 4,125 | 6,625 | 4,25
Depot | 0 | 0 | 0,5 | 0 | 0,375
andere Produkte | 0 | 0 | 0,125 | 0,5 | 1,625
- | 0 | 0 | 0 | 0 |
kein weiteres Produkt | 0,625 | 22,375 | 34,875 | 44,5 |
Kaufabstand_1
State | Kaufabstand_1 | Kaufabstand_2 | Kaufabstand_3 | Kaufabstand_4
<1 Monat | 12,875 | 10 | 6,125 | 6
1-6 Monate | 8,125 | 18 | 22 | 16,625
6-12 Monate | 5,25 | 3,125 | 3,75 | 5,5
1-3 Jahre | 11,5 | 11,375 | 10,875 | 7,375
>3 Jahre | 39,875 | 21,75 | 8,875 | 1,625
- | 22,375 | 35,75 | 48,375 | 62,875
GeburtsDatum
State | %
<1940 | 10,375
1940-1950 | 11,875
1950-1960 | 17,5
1960-1970 | 30,125
1970-1980 | 10,375
1980-1990 | 5,125
1990-2000 | 6,25
unbekannt | 8,375
GehaltsEingangDM
State | %
0 | 34
0-1000 | 2,75
1000-2000 | 3,625
2000-3000 | 9,875
3000-4000 | 8
4000-5000 | 5,125
5000-10000 | 5,625
10000-20000 | 0,625
>20000 | 0,375
unbekannt | 30
Segmente
State | %
Segment1 | 7,158
Segment2 | 15,478
Segment3 | 8,7615
Segment4 | 34,0063
Segment5 | 1,761
Segment6 | 7,7638
Segment7 | 3,6016
Segment8 | 1,0994
Segment9 | 2,0876
Segment10 | 14,3697
A separate window shows the identified segments and their size.
34% of customers fall into segment 4

Figur 2f


Produkt_1
State | Produkt 1 | Produkt 2 | Produkt 3 | Produkt 4 | Produkt 5
- 0 0 0 0 0
Giro/Gehalt 0,3679 0 0 0 0
Versicherungsprodukt 30,8751 12,1286 1,8598 0 0
Kreditkarten 0 4,4095 0 0 0
Sparen/Geldanlage 64,3461 17,6652 4,0434 0 0
Kredite 2,573 0 0 0 0
Depot 0 0 1,1025 0 0
andere Produkte 0 0 0,3676 0 0
- 0 0 0 0 0
kein weiteres Produkt 1,8379 65,7966 92,6287 100 100
Kaufabstand_1
State | Kaufabstand 1 | Kaufabstand 2 | Kaufabstand 3 | Kaufabstand 4
<1 Monat 2,2055 0 0 0
1-6 Monate 5,1451 1,1164 0 0
6-12 Monate 2,2055 0,3676 0 0
1-3 Jahre 2,9537 3,6758 0 0
>3 Jahre 21,6937 1,1111 0 0
- 65,7966 93,7292 100 100
GeburtsDatum
State | %
<1940 12,8647
1940-1950 12,4971
1950-1960 15,4377
1960-1970 23,913
1970-1980 4,4109
1980-1990 11,3949
1990-2000 15,0706
unbekannt 4,4111
GehaltsEingangDM
State | %
0 25,0156
0-1000 0
1000-2000 0
2000-3000 0
3000-4000 0
4000-5000 0
5000-10000 0
10000-20000 0
>20000 0
unbekannt 74,9844
State | %
Segment1 0
Segment2 0
Segment3 0
Segment4
Segment5 0
Segment6 0
Segment7 0
Segment8 0
Segment9 0
Segment10 0
Customers in segment 4 have no Giro account and bought only one or two products so far.
Click

Figur 2g


Produkt_1
State | Produkt 1 | Produkt 2 | Produkt 3 | Produkt 4 | Produkt 5
- | 0 | 0 | 0 | 0 | 0
Giro/Gehalt | 14,1848 | 7,2385 | 1,2043 | 0,262 |
Versicherungsprodukt | 23,5007 | 13,0306 | 10,8682 | 4,6164 |
Kreditkarten | 0 | 3,4192 | 2,7614 | 9,4669 |
Sparen/Geldanlage | 68,7839 | 22,8648 | 11,2628 | 7,3937 |
Kredite | 2,1409 | 3,6983 | 2,757 | 2,2614 |
Depot | 0 | 0 | 0,8336 | 0 |
andere Produkte | 0 | 0 | 0,2779 | 0,3898 |
- | 0 | 0 | 0 | 0 |
kein weiteres Produkt | 1,3896 | 49,7486 | 70,0348 | 75,6097 |
Kaufabstand_1
State | Kaufabstand 1 | Kaufabstand 2 | Kaufabstand 3 | Kaufabstand 4
<1 Monat | 7,1028 | 3,6861 | 4,0427 | 3,574
1-6 Monate | 5,7124 | 6,5966 | 11,5407 | 6,8657
6-12 Monate | 2,6799 | 2,2375 | 1,6852 | 2,9406
1-3 Jahre | 6,4907 | 7,8524 | 5,5526 | 3,7492
>3 Jahre | 28,2655 | 8,759 | 1,5691 | 0,8055
- | 49,7486 | 70,9684 | 75,6097 | 82,0649
One can interact with the segments in both ways:
By fixing the "GeburtsDatum" one can see which segments customers of this age group fall into.
Click
State | %
<1940 | 0
1940-1950 | 0
1950-1960 | 0
1960-1970 | 0
1970-1980 | 0
1980-1990 | 0
1990-2000 | 0
unbekannt | 0
GehaltsEingangDM
State | %
0 | 28,1942
0-1000 | 1,3968
1000-2000 | 1,6414
2000-3000 | 3,0246
3000-4000 | 3,1955
4000-5000 | 2,1691
5000-10000 | 2,1547
10000-20000 | 0,1572
>20000 | 0,0852
unbekannt | 67,9812
Segmente
State | %
Segment1 | 2,439
Segment2 | 4,8713
Segment3 | 0
Segment4 | 75,6097
Segment5 | 7,0909
Segment6 | 2,439
Segment7 | 7,56
Segment8 | 0
Segment9 | 0
Segment10 | 0